IDS 705 Final Project

Examining the Impact of Image Preprocessing on Diabetic Retinopathy Detection: A Cross-Dataset Analysis with Varied Camera Sources

Team Members:
Revanth Chowdary Ganga (rg361)
Daniela Jimenez (dj216)
Afraa Noureen (an300)
Luopeiwen (Tina) Yi (ly178)

Team Number: 9
Github Repository: https://github.com/nogibjj/Diabetic-Retinopathy

Abstract

Diabetic retinopathy (DR) is a major health concern worldwide, with a significant impact in low- and middle-income countries (Vujosevic et al., 2020). Machine learning models show promise in aiding diagnosis (Oh et al., 2021), but their performance may vary across different datasets due to disparities in image acquisition. This project explores the impact of image preprocessing on enhancing DR detection models across various camera sources. By using datasets with distinct characteristics and applying processing techniques such as image cropping, luminosity adjustment, and color normalization, we aimed to improve model generalization. Our experiments involved using EfficientNetB0 (Tan & Le, 2019), an existing pre-trained Convolutional Neural Network (CNN) architecture, as part of our effort to replicate the paper titled "Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning" (Alyoubi et al., 2021). Our goal was to study the impact of different image preprocessing options on the generalizability of the models and to identify a suitable combination of these options that results in the best generalized performance of the models. Preliminary results from our base paper (Alyoubi et al., 2021) suggest that image preprocessing can reduce performance disparities across datasets without significantly compromising overall model effectiveness. This project aims to contribute to the understanding of the effect of different image preprocessing methods on enhancing DR diagnosis, thereby helping mitigate the burden of this sight-threatening complication of diabetes.

Introduction

Diabetic Retinopathy is the leading cause of blindness among individuals aged 20 to 74 years and is one of the foremost complications of diabetes, as well as a major cause of preventable blindness (Lee et al., 2015). Diabetes, affecting 422 million people worldwide, is on the rise, particularly in middle- and low-income countries (World Health Organization, 2019). It is crucial, therefore, that the health systems in these countries are equipped with the necessary tools and technologies to manage this chronic disease's outcomes. Machine learning models introduce new approaches to medical diagnosis and have the potential to relieve overburdened health services by preventing the diseases that strain them the most. For us data scientists, it is gratifying to apply our expertise to address issues that not only present global challenges but also impact our own nations.

Machine learning techniques have offered alternative methods for diagnosing Diabetic Retinopathy by analyzing retinal images (Pires et al., 2019). The cameras used to capture these images vary in specifications, including color profiles and field of view (FOV). We hypothesized that a model trained on images from one camera type may underperform when presented with images from another camera. The core issue we addressed is whether the generalizability of detection models can be enhanced through preprocessing techniques like luminosity adjustment, cropping, and color normalization. The results from our study suggest that using certain preprocessing techniques on retinal images could improve both the generalizability of the models and their effectiveness in the accurate diagnosis of diabetic retinopathy, regardless of the camera type.

Background

Diabetic Retinopathy (DR), a common complication associated with diabetes (Zochodne & Toth, 2014b), results in lesions on the retina that impair vision. It is essential to detect DR early to avoid permanent blindness (Kansu, 2010b), as the condition's damage cannot be reversed and treatments (Ascher, 2002b) focus only on maintaining existing vision. Timely detection and treatment significantly reduce the risk of vision loss (Güler et al., 2023b). Recently, deep learning (Kotu & Deshpande, 2019b) has proven to be a highly effective method, demonstrating exceptional results in many areas, especially in the analysis and classification of medical images (Cinaglia et al., 2019b). Convolutional Neural Networks (CNNs) (Dhanya et al., 2022b), a type of deep learning technique, have become particularly prominent for their effectiveness in evaluating medical images.

The 2020 ScienceDirect publication “Diabetic retinopathy detection through deep learning techniques: A review” (Alyoubi et al., 2020) analyzed the use of deep learning (DL) in detecting and classifying diabetic retinopathy across 33 studies. These studies focused on leveraging DL for DR screening systems to address the rising need due to the growing number of diabetic patients. A significant advantage of DL in DR detection is its ability to bypass the need for manually selected features required for traditional machine learning, although it demands substantial data for training. To combat overfitting and data scarcity, most research utilized data augmentation and public datasets; 94% used public datasets, with 59% combining multiple sources to enrich data size and diversity for better DL model evaluation.

A notable challenge highlighted was the large data size needed for effective DL training, emphasizing the necessity for more extensive or refined datasets to improve DL system accuracy. Regarding DL techniques, there's a variation between studies constructing their own convolutional neural network (CNN) architectures and those opting for pre-existing models like VGG, ResNet, or AlexNet through transfer learning. Custom-built CNNs tended to show higher accuracy but required more effort and time, suggesting a need for further research to compare these approaches. The review also found that the majority of studies (73%) classified images simply into DR or non-DR categories, with only 27% classifying into specific DR stages and 30% identifying affected lesions. This indicates a research gap in developing reliable DR screening systems that can accurately determine DR stages and lesion types, which is crucial for effective patient monitoring and treatment.

We replicated EfficientNet B0, a pre-trained CNN model, as outlined in the 2021 Sensors Journal research paper "Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning" (Alyoubi et al., 2021). The CNN models described in this paper are designed to classify the stages of DR, incorporating preprocessing techniques such as Image Enhancement, Noise Removal, Cropping, Color Normalization, and Data Augmentation. These models also adjust their architecture to align with other effective works. The resulting model, EfficientNetB0, achieved an accuracy of around 82% on both the Asia Pacific Tele-Ophthalmology Society 2019 (APTOS) and Dataset for Diabetic Retinopathy (DDR) datasets. Despite these high-performance results, the research does not address the effects of utilizing different data sources or the impact of various image preprocessing methods, whether applied individually or in combination, on a single model’s

performance. By replicating the model and employing different data sources and preprocessing techniques, we aimed to test our hypothesis that image pre-processing options can result in a more generalized performance of DR detection models.

Data

One crucial aspect of our research involved using images captured by different cameras to establish a baseline performance for the models and to evaluate the enhancements in performance following image preprocessing. While there were many sources of retinal images available for DR analysis, such as those on Kaggle (*Diabetic Retinopathy Detection | Kaggle*, n.d.), these sources often comprised images from multiple origins and cameras, which presented a high risk of data leakage if we compared one dataset with another. After a thorough investigation for datasets that explicitly documented the camera used and the specifications of the images, we selected two datasets for our study: Messidor2 (Maffre, 2022b), referred to hereafter as Messidor, and IDRID (Porwal, 2019b). These datasets were chosen for their clarity in camera specification and their suitability for rigorous analysis.

DDR

To verify that the construction of the EfficientNetB0 model architecture was identical or as close as possible to the one used in the research paper (Alyoubi et al., 2021), we used the same dataset that was used in the paper to validate the model construction and fine-tune the hyperparameters. The dataset used was DDR (Li et al., 2019), collected from 2016 to 2018 from 147 hospitals in China. The images were captured using a variety of cameras and included 13,673 eye fundus images with image annotation and classification available. This dataset is graded on a scale from 0 to 4 with the majority falling in the normal category (Figure 1). Using DDR for hyper-parameter tuning also ensures that the model is not influenced or biased towards any of the datasets used in the experiment.

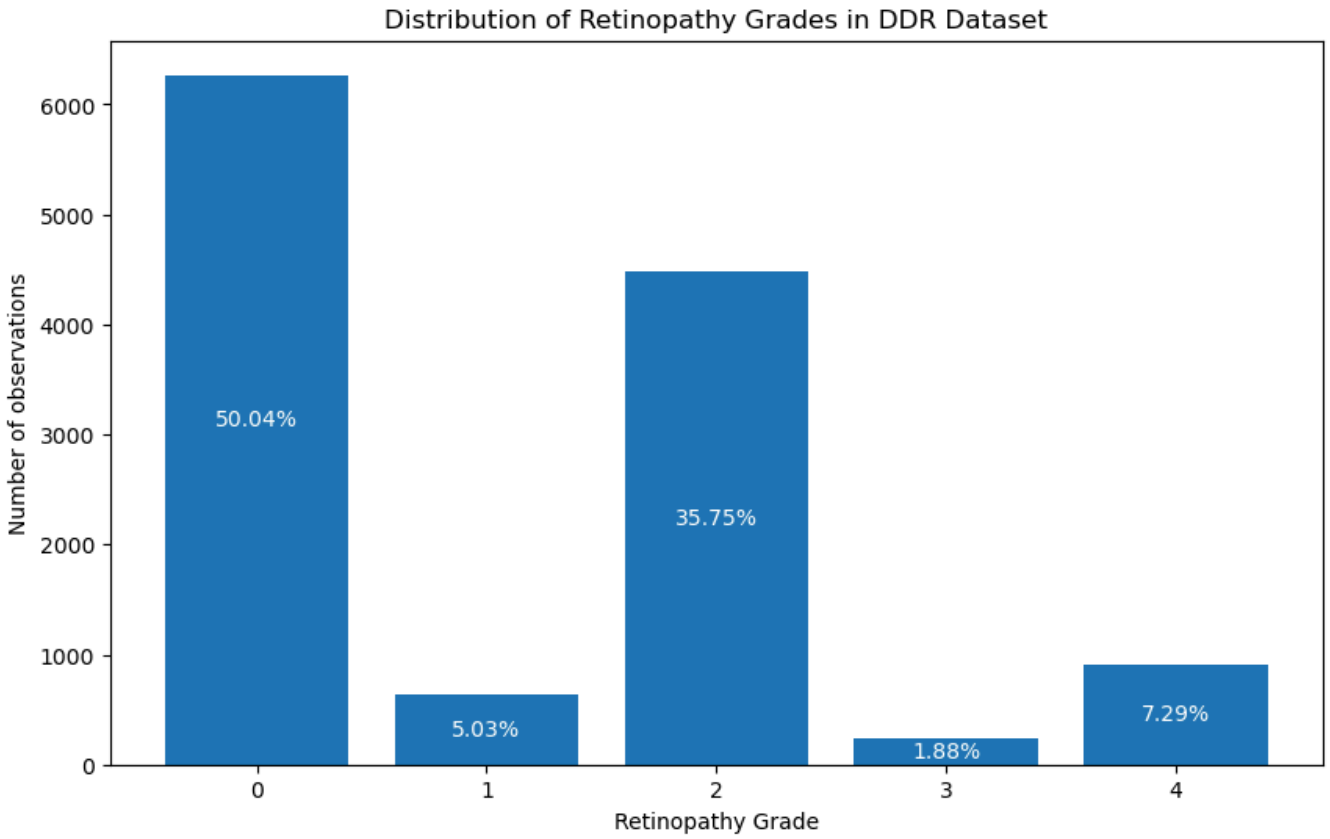

Figure 1. Distribution of Retinopathy Grades in DDR Dataset

Messidor

The Messidor dataset comprises 1,200 eye fundus color images of the posterior pole, sourced from three different ophthalmologic departments. All images were captured using a Topcon TRC NW6 camera with a 45-degree field of view. The images are in one of the following resolutions: 1440x960, 2240x1488, or 2304x1536, depending on the ophthalmologic department from which they were obtained.

Out of the 1,200 images, there are a few duplicates and some that are incorrectly marked; the details of these were provided by the dataset source and were corrected accordingly. Diabetic Retinopathy in the Messidor dataset is graded on a scale from 0 to 3 (best to worst), with the majority of the observations categorized as "0" (i.e., Normal) (Figure 2.).

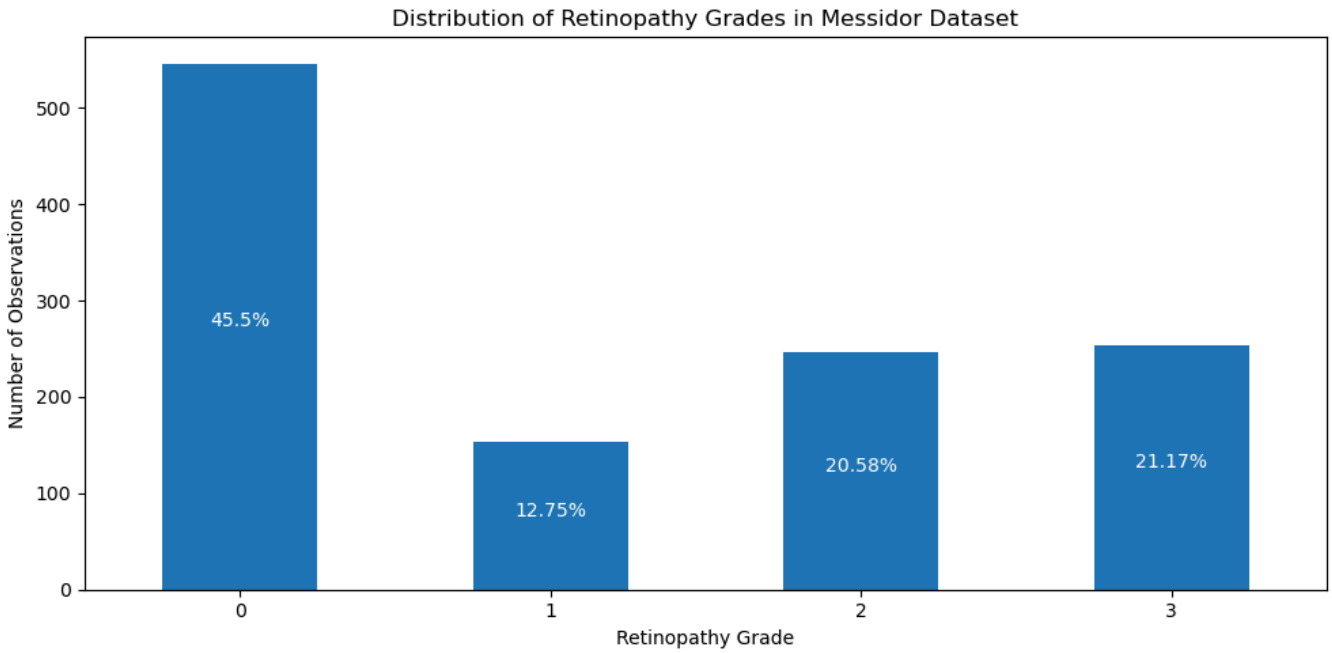

Figure 2. Distribution of Retinopathy Grades in Messidor Dataset

IDRID

The IDRID dataset comprises 516 images captured using a Kowa VX-10α camera with a 50-degree field of view. Unlike Messidor, all images in the IDRID dataset have an uniform resolution of 4288×2848 pixels.

The data source confirmed the absence of duplicates or mislabeled images within the dataset. Diabetic Retinopathy is initially graded on a scale of 0-4 in this dataset, aligning with international standards. However, since the Messidor dataset uses a grading scale of 0-3, the images in IDRID were re-labeled to harmonize with Messidor's scale to facilitate effective cross-testing (refer to Appendix 1 for details). Further information on the image reclassification process will be discussed in the preprocessing part of the experiment section. Post the relabeling, the majority of the observations in the IDRID dataset were from category 2 (Figure 3.).

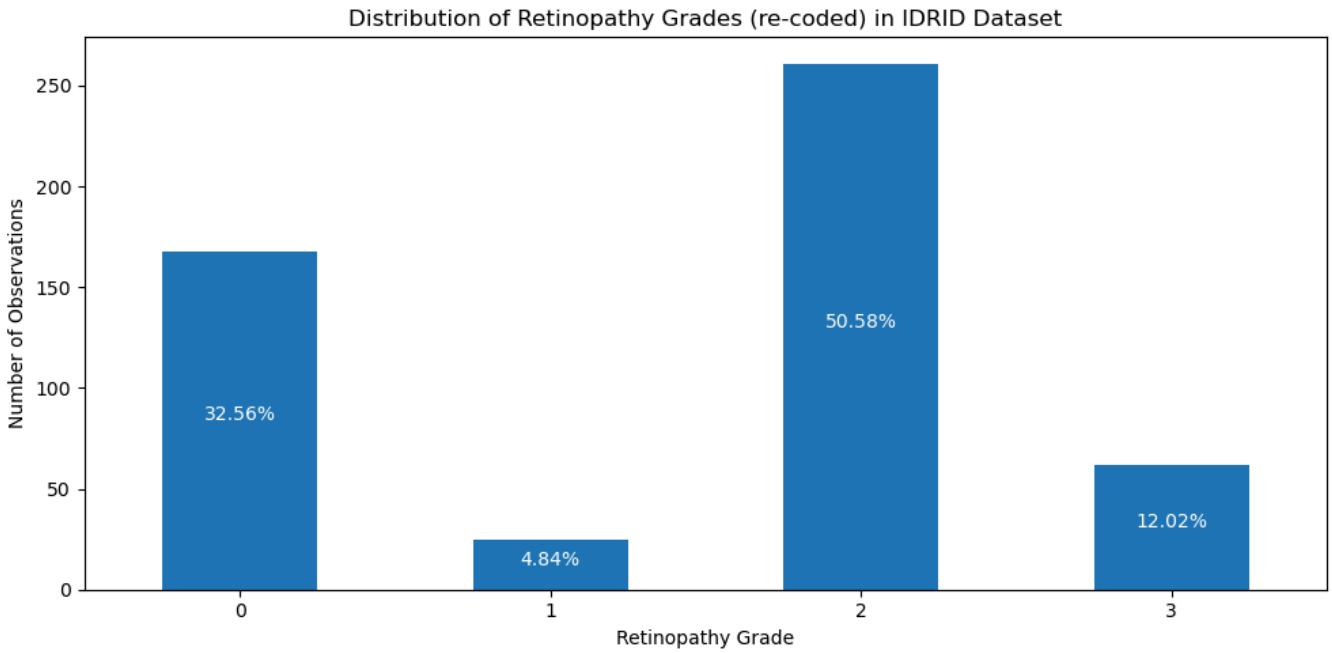

Figure 3. Distribution of Retinopathy Grades (re-coded) in IDRID Dataset

Our dataset selection aims to create a robust foundation for studying the impact of different preprocessing techniques on DR diagnosis to improve the model's generalization performance across different fields of view and images taken from different sources (Figure 4.).

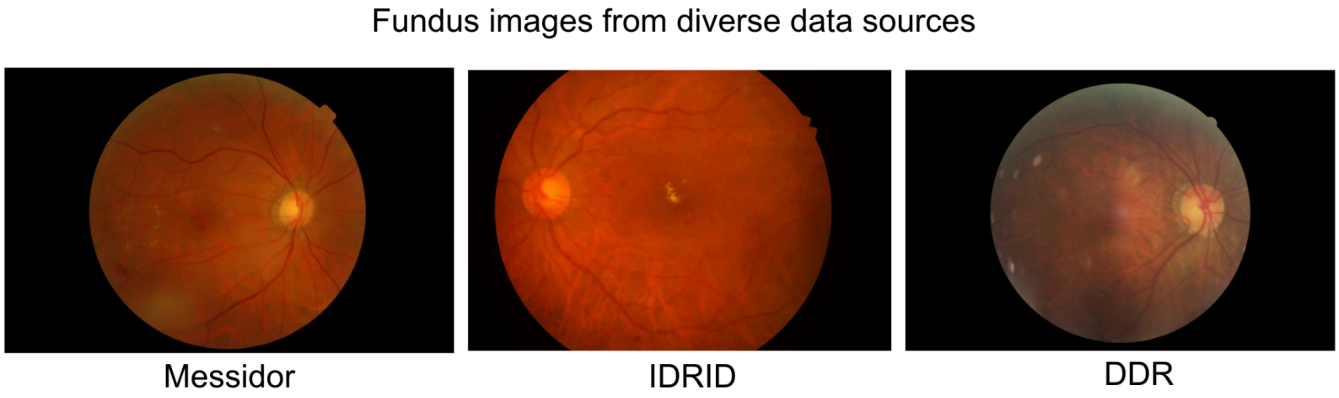

Figure 4. Examples of fundus images from the different data sources: Messidor, IDRID, DDR

Experiments

Part 1: Methodology

In this section, we outline the experiments conducted to investigate how image preprocessing influences the performance of diabetic retinopathy detection models. We structured our approach into three distinct workflows: 1) Validating the model architecture using the DDR dataset, 2) Establishing a baseline model

using raw images from the Messidor and IDRID datasets, and 3) Evaluating various preprocessing methods, both individually and in combination, to identify the most effective technique. We replicated the base paper's approach by using a train/test split without a separate validation dataset in our experiments since our primary goal was to compare the model performances. Through systematic testing at each stage, our goal is to identify the best methods for enhancing model performance.

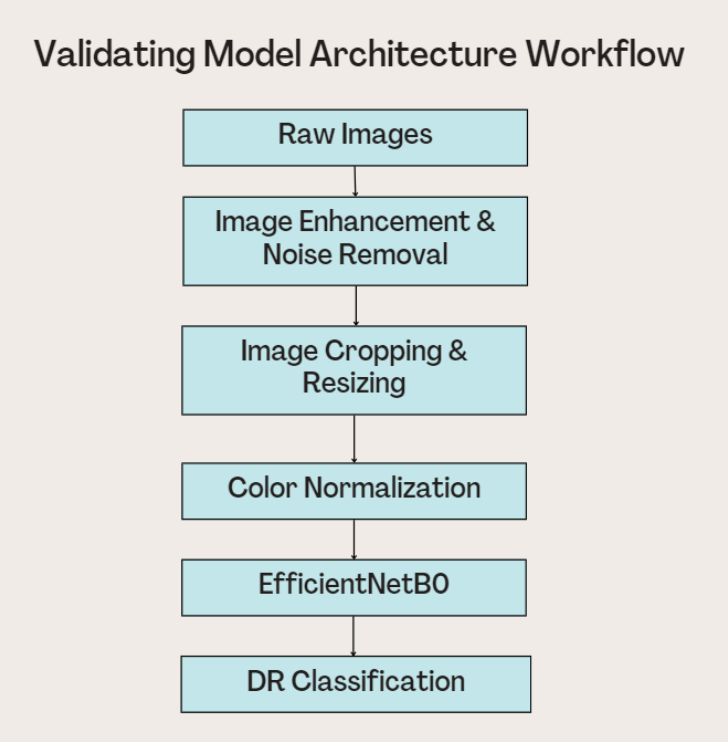

Figure 5. Validating Model Architecture Pipeline

As illustrated in Figure 5, we established a validation model pipeline to replicate the model used in the original study (Alyoubi et al., 2021). This process began with raw retinal images from the DDR dataset, which were first subjected to image enhancement and noise removal techniques to enhance quality and clarity. These images were then cropped to eliminate irrelevant black pixels and focus on the region of interest, followed by resizing to a uniform dimension of 224x224, optimizing them for input into our model. Subsequently, color normalization was applied to standardize pixel intensities across different images. After these preprocessing steps, the images were directly fed into the EfficientNetB0 model for DR classification. This workflow was essential to validate our model architecture, ensuring that our validation results closely align with those presented in the original paper before proceeding with our proposed experiments. Notably, we did not employ the data augmentation process described in the original studies in our validation model, as the DDR dataset contains 13,673 images, and with data augmentation, the number would have swelled to around 275,000 images, exceeding our computational resources even with GPU support. Given the substantial size of the original dataset, we anticipate that excluding the data augmentation step does not critically impact the performance of the validation model. The validation model helped us confirm our preprocessing mechanism and model architecture were correct before we proceeded to the subsequent experimental stages. Further information on the validation of our model architecture is discussed in the subsequent results section.

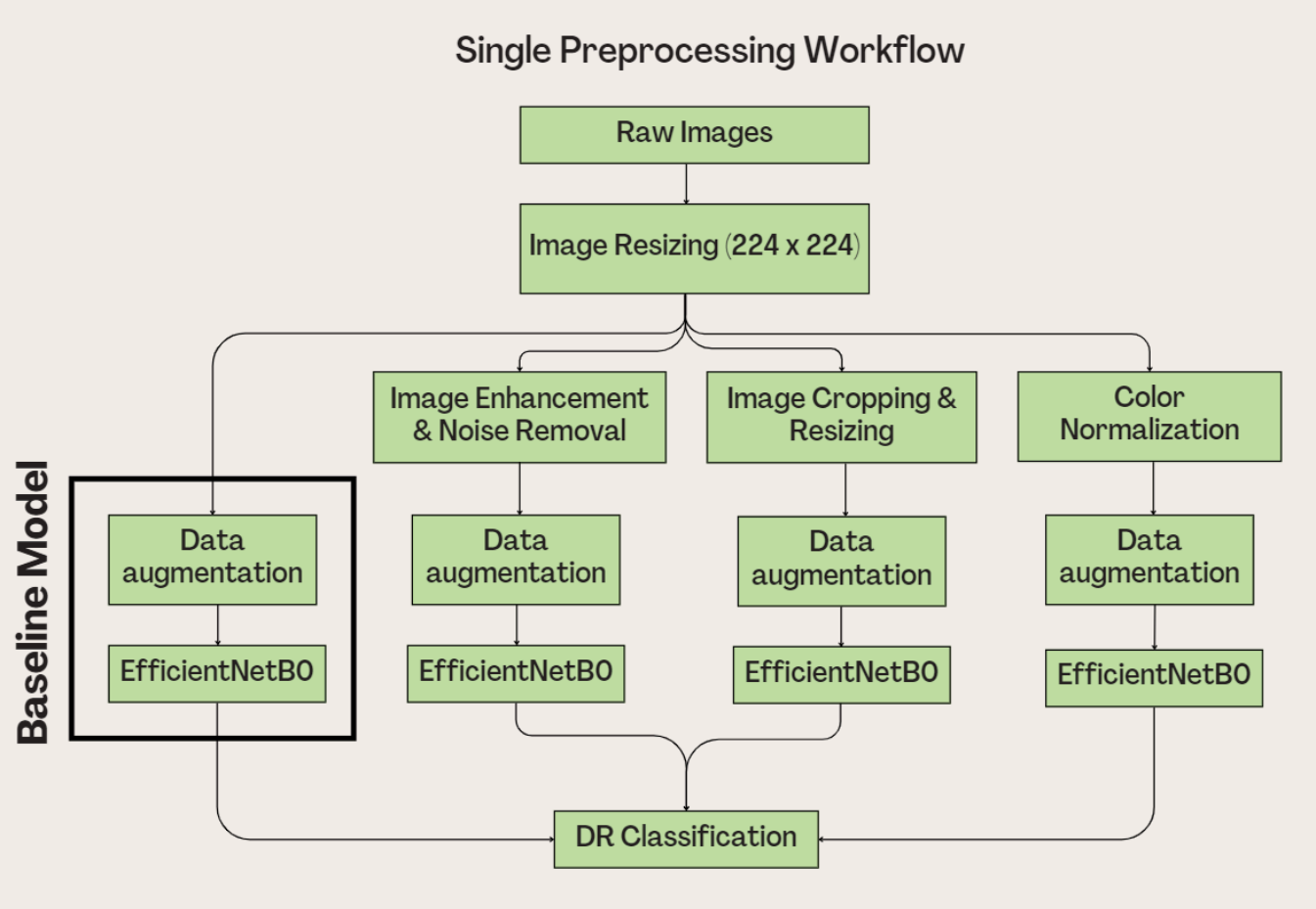

Figure 6. Single Preprocessing Methods Pipeline

After validating our model architecture, we proceeded with the single preprocessing methods pipeline for each dataset (Messidor & IDRID) in our proposed experimental workflow, as shown in Figure 6. This pipeline began with the basic image preparation, where raw retinal images were resized to standard dimensions of 224x224 to ensure consistency across the dataset and facilitate model compatibility. The resized images were then processed through four parallel pipelines.

The first pipeline served as our baseline model, involving the direct usage of resized images. This setup established the baseline model as a scenario of control for our experiments. The second pipeline utilized image enhancement and noise removal to improve image quality. The third pipeline focused on image cropping and resizing to concentrate on the region of interest. The fourth pipeline involved applying color normalization to standardize pixel intensities across images. Subsequently, the processed images from each pipeline underwent various forms of data augmentation. Finally, these augmented images were fed into the model (EfficientNetB0) for classification. Each image preprocessing step in this workflow is detailed in the following preprocessing subsection.

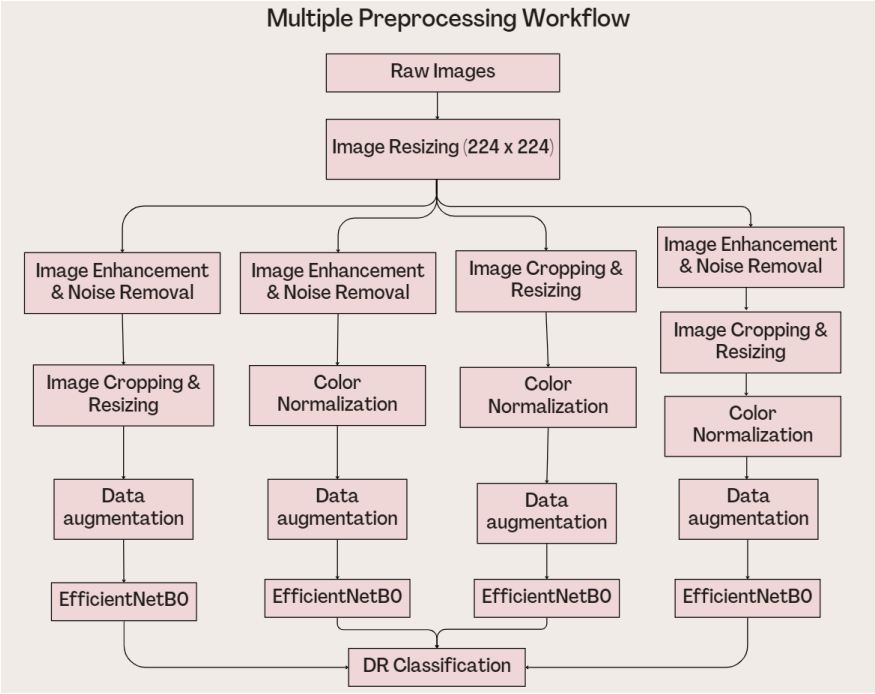

Figure 7. Multiple Preprocessing Methods Pipeline

In addition to the single preprocessing methods pipeline, we also implemented a multiple preprocessing methods pipeline as part of our proposed experimental workflow, as depicted in Figure 7. In this multi-preprocessing pipeline, after resizing the raw images, we applied different combinations of two out of the three preprocessing methods, and eventually combined all preprocessing methods before proceeding to data augmentation and then to the EfficientNetB0 for final DR classification. This strategy helped us assess the impact of different preprocessing combinations on the performance of our diabetic retinopathy classification model.

Part 2: Preprocessing

(1) Image Re-classification

As mentioned earlier, the images in the Messidor dataset are classified on a scale of 0-3. In this scale, two of the standard categories—Moderate and Severe—were combined into a single category (stage 2). To maintain consistency in our experiments, it was necessary to ensure that both datasets (Messidor and IDRID) used the same classification scale. Since we lacked the required information to split the combined category in Messidor, we merged the corresponding categories in IDRID to align with the Messidor classification scale, based on publicly available information about the datasets (as shown in Table 1.) (refer to Appendix 1 for details).

S.No	DR Stage	Details	Messidor	IDRID
1	Healthy / No DR	No Abnormalities	0	0
2	Mild DR	Only Microaneurysms (MA)	1	1
3	Moderate DR	More than just MA but less than severe DR	2	2
4	Severe DR	Venorous beading and Intraretinal Microvascular Abnormalities	2	3
5	Proliferative DR	Pre-Retinal Hemorrhage	3	4

Table 1. Image Re-classification Preprocessing

(2) Image Resizing

The Messidor Dataset comprises images of three different sizes: 1440x960, 2240x1488, and 2304x1536, while the IDRID dataset contains images of size 4288x2848. All images were resized to a common resolution of 224x224 (as shown in Figure 8.). Although the initial resolutions varied, each had a 3:2 ratio. Consequently, resizing them to the same dimensions meant that the aspect ratio change was consistent across all images.

The images were resized to a smaller resolution for several reasons:

1. **Consistency**: A model trained on a specific resolution will have tailored parameter values and may not perform optimally if presented with images of a different resolution.
2. **Resource Efficiency**: Reducing the image size decreases the amount of storage and computation required.
3. **Model Compatibility**: The model we used, EfficientNetB0, was optimized for an image size of 224x224 in the original study (Alyoubi et al., 2021). Therefore, we needed to resize the images to this resolution to utilize the model effectively.

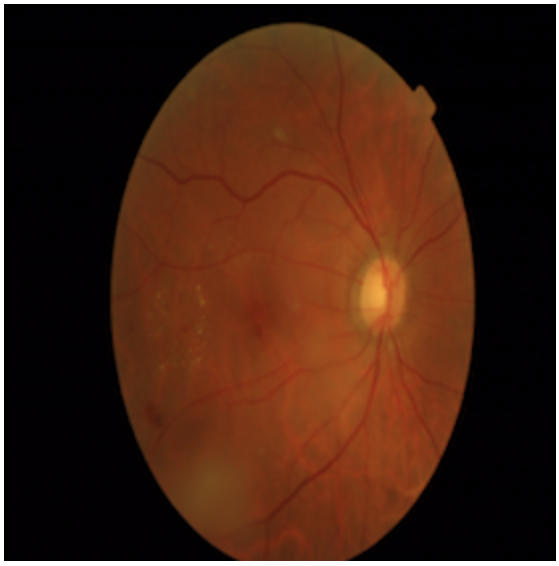

Figure 8. Resized Messidor Image

(3) Test Dataset Creation

The distribution of the classes was different - both between the datasets and within the different sources in Messidor. To ensure proper comparison when we were evaluating the models, we wanted to have the 2 test datasets (one from each dataset) identical to each other in terms of the distribution. Additionally, since we wanted to study the performance of the model in classifying the stages of DR instead of having just a binary classification, we decided to have the same number of samples from each category which eliminates any bias towards a single dataset along with helping us evaluate the models ability to distinguish the different stages.

We had the following constraints or requirements from the 2 test datasets:

- They should be of the same size
- They should have equal number of samples from each category
- They should represent all of their sources equally (applicable to Messidor)
- The number of test samples should not be too high such that we don't have enough training samples (applies to IDRID category 1)

Based on these requirements, we decided to keep the test size as a fixed value of 24 images (6 per category) from each dataset.

For Messidor, the test dataset was created using the following approach:

- The Dataset was shuffled once to break any natural orders.
- The samples were taken by accounting for both; the category and the source.
- 2 samples were taken from each of the unique source-category combinations, resulting in 24 samples in total.

For IDRID, since all of the images were from the same source, the test dataset was created by taking 6 samples from each of the 4 (recoded) categories resulting in 24 samples in total.

(4) Image Enhancement and Noise Removal (IENR)

We applied the following image preprocessing to both the train and test datasets:

Image Enhancement: Two primary techniques were used. The first technique, known as the enhancement of luminosity (Foracchia et al., 2005), aims to increase the clarity and visibility of the images. The second, Contrast Limited Adaptive Histogram Equalization (CLAHE), is renowned for its efficacy in augmenting the contrast of these images. Specifically, CLAHE is adept at improving the low-contrast areas within medical images, making it a valuable tool in medical diagnostics. To implement CLAHE, we transformed the images to the LAB color space, focusing on the L channel that represents lightness. Given that retinal images inherently possess higher contrast, they were particularly suitable for this method. We then applied CLAHE with a tile size of 8×8 and a clip limit of 5.0.

Image Noise Removal: To counteract the introduction of noise into the images from the CLAHE method, noise removal is executed using a Gaussian filter, described mathematically as

$$G_0(x, y) = Ae^{\frac{-(x-\mu_x)^2}{2\sigma_x^2} + \frac{-(y-\mu_y)^2}{2\sigma_y^2}}$$

Where μ denotes the mean, A the amplitude, and σ the standard deviation for the variables x and y.

This process ensures that the enhanced images retain their improved clarity and contrast while minimizing any added noise, thus providing a clearer and more detailed view for medical analysis (as shown in Figure 9.).

Figure 9. IENR preprocessing on Messidor Image

(5) Image Cropping (Crop)

We performed cropping on the images to remove superfluous black pixels surrounding the retina. This process, as illustrated in Figure 10, was crucial in focusing the model's attention on the relevant features within the retinal images.

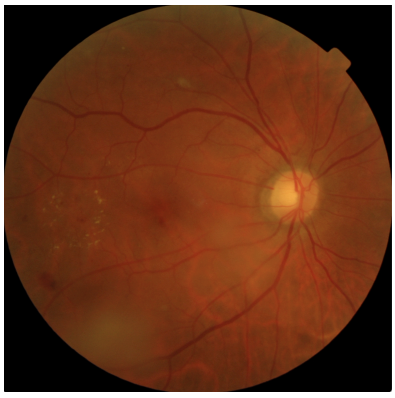

Figure 10. Cropping preprocessing on Messidor Image

(6) Normalization by Color (CN)

Color normalization plays a crucial role in the preprocessing of retina images, especially given the diverse conditions under which these images are captured. Factors such as the patient's age, ethnicity, and varying lighting conditions during image capture can introduce significant variations in pixel intensity across images, leading to inconsistencies that could hinder analysis.

To address these challenges, a normalization process was applied to **each channel** of the RGB images. This process involved adjusting the images based on their mean and variance. For every x pixel value in the training RGB retina images, the following normalization formula was used:

$$z = \frac{(x - u)}{s}$$

, where u represents the mean and s the standard deviation of the training RGB retina images. It's important to note that this normalization was conducted separately for each color channel (Red, Green, and Blue).

This normalization technique effectively standardizes the pixel intensity values across different images, thereby reducing unnecessary variation and making the images more comparable and analyzable across diverse conditions (as shown in Figure 11).

Figure 11. Color Normalization preprocessing on Messidor Image

(7) Data Augmentation

To enhance the training dataset and bolster the generalization and performance of our model, data augmentation techniques were implemented. Specifically, data augmentation was performed on the IDRID and Messidor datasets to increase the number of samples. This strategy involved dynamically modifying the images through various transformations to create a more diverse set of training data. The transformations included rotation, flipping, shearing, and translation, along with adjustments to the

images' brightness levels, either darkening or brightening them randomly. This process not only enriched the dataset but also introduced a level of robustness to the model, enabling it to better handle a wide range of input variations and improve its predictive accuracy. (Refer Appendix 2. for details)

Part 3: Modeling

(1) EfficientNetB0

We utilized transfer learning with EfficientNetB0 (Tan & Le, 2019), a pre-trained model renowned for its exceptional performance on the ImageNet dataset. We used two transfer learning strategies: feature extraction using EfficientNetB0 and fine-tuning EfficientNetB0 by initializing its weights with those from ImageNet and retraining it using our own datasets. To adapt EfficientNetB0 further to our needs, its top layers were discarded and substituted with a new configuration: a Global Average Pooling (GAP) layer, two fully connected (FC) layers, and a SoftMax layer (as shown in Figure 12.). Additionally, to address the challenge of overfitting, a Dropout mechanism with a rate of 0.5 was integrated into the FC layers. Our decision was guided by experiments comparing performance with and without Dropout. We found that including Dropout consistently improved model performance, so we used it in our final configuration. The input size used is 224 × 224 × 3 for EfficientNetB0.

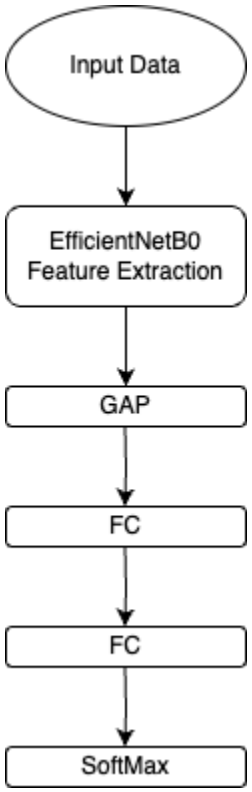

Figure 12. EfficientNetB0 Model Architecture

(2) Hyperparameter Configuration

The hyperparameter settings for this model were carefully chosen to help it learn effectively and prevent overfitting. We used the Stochastic Gradient Descent (SGD) optimizer which helped in making smooth adjustments during training. For EfficientNetB0, the maximum learning rate was set to enable the model to adjust its learning speed as needed. The base learning rate provided a stable starting point for learning. The 'triangular' mode in the learning rate scheduler allowed the model to adapt its learning rates effectively. Automatic adjustment of class weights ensured balanced attention to different data types. Data augmentation, applied multiple times, generated varied training data to boost model robustness. Additionally, a dropout mechanism was used to prevent overfitting by randomly deactivating parts of the model during training, encouraging better generalization. These settings collectively helped optimize the model's learning process, improve adaptability to diverse data, and reduce the risk of overfitting. (Refer Appendix 3. for details)

Part 4: Evaluation Method

Accuracy was chosen as a primary metric in this paper due to its simplicity and ease of interpretation, especially for a multi-class classification problem like diabetic retinopathy detection & classification. DR classification involves multiple stages or severity levels, ranging from mild to severe, making it suitable for a multi-class approach. Since our primary goal was to only compare the relative performance between the models, accuracy was sufficient to make this comparison. We also explored unconventional methods to evaluate the model performance which has been detailed further in the results section.

While many papers focus on binary classification (detecting the presence or absence of DR), our aim was to develop a model that can provide more detailed information about the severity of DR. This finer-grained approach is valuable in medical contexts, as it allows for better patient management and treatment planning. By using a multi-class classification, our model can provide clinicians with more nuanced insights into the progression of the disease, which can guide appropriate interventions.

To establish a baseline, we trained and tested the models using raw-resized color images. Once the baseline performance was determined, we proceeded to preprocess the images—adjusting luminosity, cropping, and normalizing, among other techniques. These processed images were then used to train and test the models, allowing us to assess the impact of various preprocessing methods on model performance.

For each type of model that we evaluated, we used the following approach:

1. For each type of model evaluation, one model was trained on Messidor and the other on IDRID data, then they were evaluated using both of the test datasets separately.
2. Once the performance of the model on the Raw (only resized) images had been established, the models were evaluated on the images that underwent one type of processing, to see if that particular modification resulted in any improvements

3. Later, the performance of the model was evaluated on different combinations of the pre-processing methods and finally with all of them combined. This helped to have a thorough understanding of how these preprocessing techniques impacted the performance of the model both when used by themselves and in combination with other pre-processing techniques

For the sake of simplicity, we will be using the following terms when we talk about the performance of the model:

1. Self-accuracy: is the accuracy obtained when the test data is from the same dataset as the training data i.e. (trained and tested on messidor or trained and tested on IDRID)
2. Cross-accuracy: is the accuracy obtained when the test data is from a different dataset as compared to the training data i.e. (trained on Messidor and tested on IDRID or trained on IDRID and tested on Messidor)

If a dataset name is used before these terms, that indicates the training data, e.g.:
- Messidor self-accuracy: Trained on Messidor and Tested on Messidor
- Messidor cross-accuracy: Trained on Messidor and Tested on IDRID

Results

Model Architecture Validation Result

Our architecture validation model (Figure 5), trained on the DDR dataset without data augmentation, achieved an accuracy of 81.38%, which is very close to the 82.2% reported in the original study that employed the same model pipeline but included data augmentation. The slight discrepancy in results could be attributed to randomness in the train-test split and the absence of data augmentation in our validation approach. Consequently, our model was successfully validated and prepared for use in the subsequent experimental stages.

Experiment Results

Table 2 presents the self-accuracy, cross-accuracy, and absolute difference in performance between the datasets for each of the experiments conducted. As discussed previously, we established a baseline using models trained on resized images from both Messidor and IDRID datasets, without applying any of the preprocessing methods under investigation. For the experimental section, we applied the image preprocessing methods both individually and in combination across different model pipelines, training on Messidor and IDRID individually.

			Accuracy On		
	Processing	Trained On	Messidor	IDRID	Abs. Difference
RAW	Resizing	Messidor	58.33%	45.83%	12.50%
		IDRID	20.83%	62.50%	41.67%
Single Preprocessing	IENR	Messidor	58.33%	41.67%	16.67%
		IDRID	29.17%	54.17%	25.00%
	Crop	Messidor	66.67%	45.83%	20.83%
		IDRID	29.17%	66.67%	37.50%
	CN	Messidor	58.33%	45.83%	12.50%
		IDRID	29.17%	58.33%	29.17%
Multiple Preprocessing	IENR + Crop	Messidor	54.17%	41.67%	12.50%
		IDRID	20.83%	41.67%	20.83%
	IENR + CN	Messidor	50.00%	37.50%	12.50%
		IDRID	25.00%	58.33%	33.33%
	Crop + CN	Messidor	70.83%	41.67%	29.17%
		IDRID	33.33%	62.50%	29.17%
All	All	Messidor	62.50%	37.50%	25.00%
		IDRID	29.17%	58.33%	29.17%

Table 2. Accuracy Results of Baseline Model and Experiments

Figure 13. shows the visualization of the performance of models trained on the Messidor dataset with different preprocessing methods.

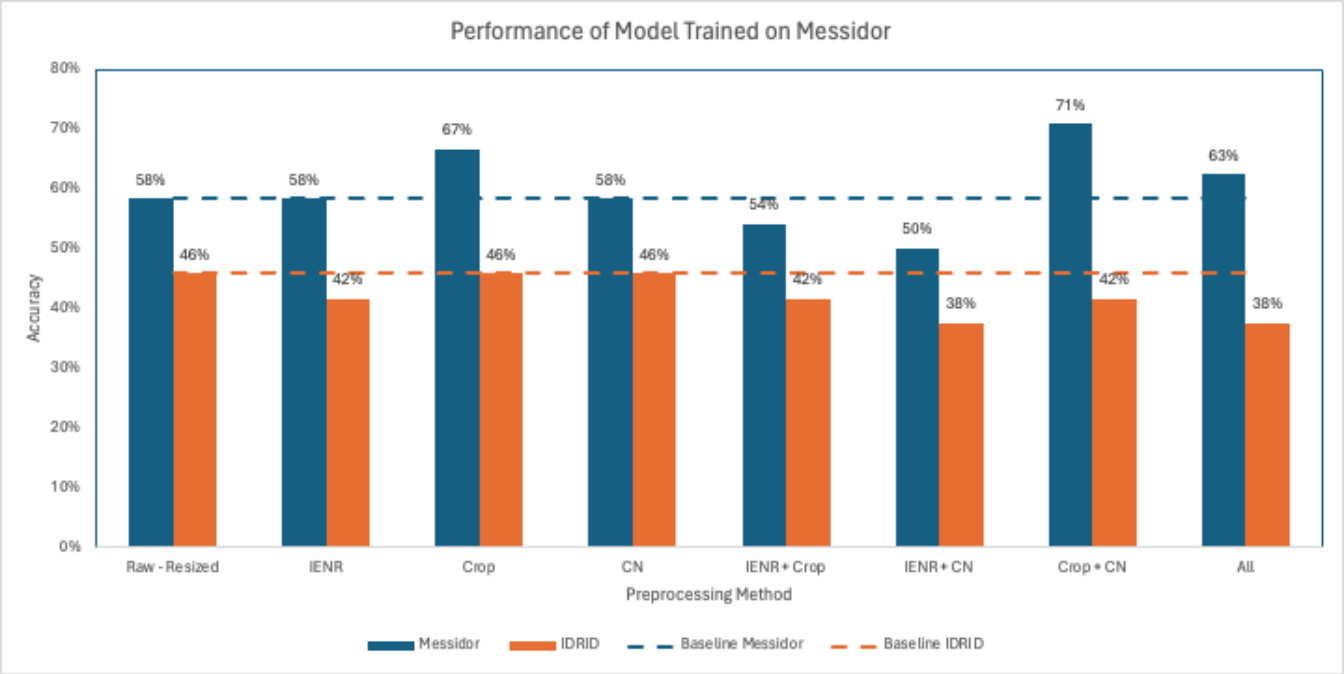

Figure 13. Performance of Models Trained on Messidor

Figure 14. shows the visualization of the performance of models trained on the IDRID dataset with different preprocessing methods.

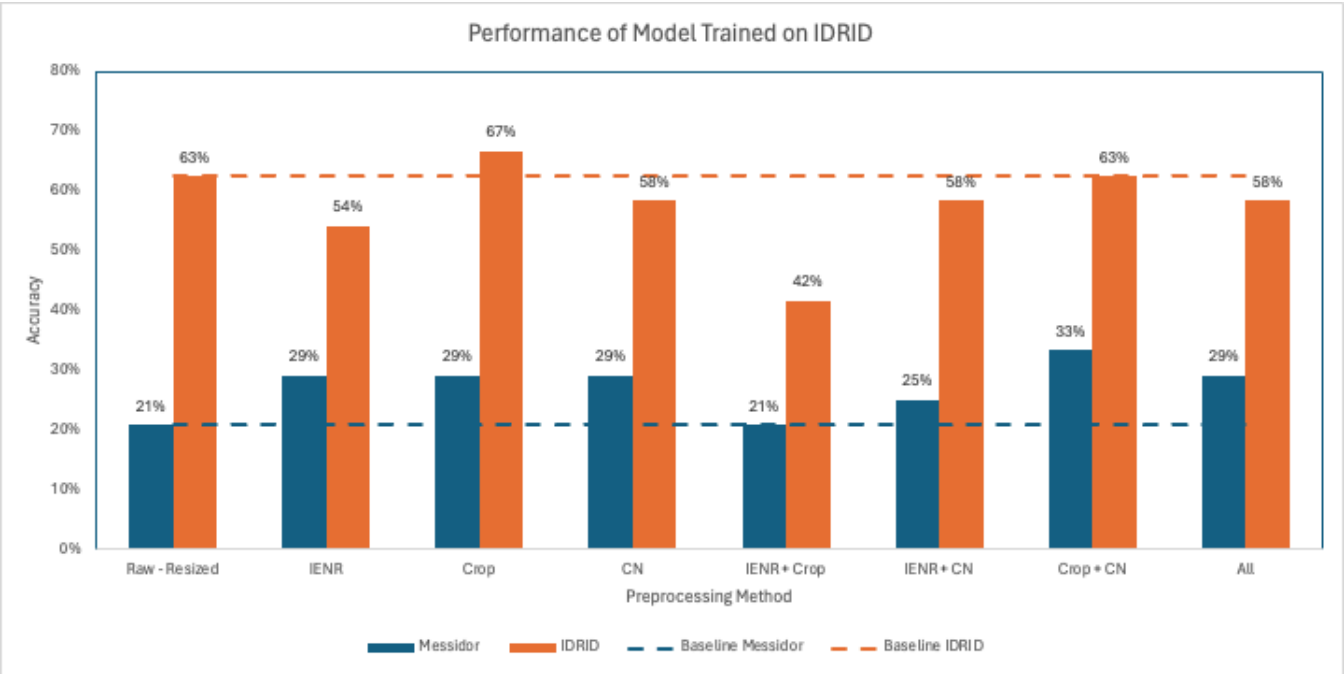

Figure 14. Performance of Models Trained on IDRID

Baseline Model Analysis

As observed in Table 2, the self-accuracy in all the experiments was higher than the cross-accuracy for the respective experiments. This analysis of the raw data provides several crucial insights: it confirms our hypothesis, validates the reliability of both datasets (indicating that the lower accuracy on either dataset is not due to poor dataset quality), and establishes a baseline for subsequent experiments.

Model with Preprocessing Analysis

With the hypothesis confirmed and the baseline for comparison established, we proceeded to analyze the impact of different preprocessing techniques on the model's performance. The analysis was conducted in stages: first by testing each image preprocessing method individually, then in combinations of two, and finally using all the methods together. The observations were as follows:

(1) Image Enhancement and Noise Removal (IENR)

In the initial stages, we anticipated that IENR would be the most efficient technique; however, it proved to be the least beneficial and often even impaired performance, as shown in Figures 13 and 14. On comparison with the baseline, the results indicate that for models trained on Messidor, self-accuracy either remained stable or decreased while cross-accuracy decreased. This led to an increased absolute difference, which is contrary to our objectives. For models trained on IDRID, although the reduction in absolute difference initially seemed favorable, it was at the cost of diminished performance on the IDRID dataset itself, which is not a desirable outcome. We speculate that the ineffectiveness of IENR might be due to its potential to "smooth out" crucial features such as micro-aneurysms and other eye abnormalities crucial for differentiating DR stages. Given that our approach involved multi-class rather than binary classification, and considering that key features might have been obscured, the model may struggle to differentiate between classes, leading to suboptimal performance. Additionally, we observed that the IENR process might interfere with the cropping process. IENR added values to the padding pixels, which resulted in less effective cropping compared to using raw images. This interference could also explain why models incorporating IENR exhibited poor performance, possibly due to the negation of the effects of cropping and color normalization.

(2) Cropping (Crop)

As shown in Figures 13 and 14, the performance outcomes for cropping aligned with our expectations; they generally mirrored the results of the raw data or showed an improvement. This likely stems from the fact that cropping removes the non-informative "black" padding pixels, which do not contribute to the model's predictive capabilities. By eliminating these extraneous pixels, we either maintain consistent performance or, in some cases, enhance it by directing the model's focus towards the relevant features of the image.

(3) Color Normalization (CN)

As illustrated in Figures 13 and 14, the implementation of color normalization as a standalone preprocessing method did not yield improvements for the Messidor dataset; however, for IDRID, despite a slight decline in self-accuracy, cross-accuracy increased significantly, thus reducing the absolute difference. The benefits of color normalization became more evident when combined with cropping, leading to the most favorable overall outcomes. We hypothesize that color normalization alone can potentially reduce the disparities between the color profiles captured by different cameras. However, its effectiveness may be hindered by the presence of black padding pixels that interfere with the process. In contrast, when cropping is applied before color normalization, it significantly reduces these padding pixels, thereby enhancing the effectiveness of color normalization on both datasets.

Additional Analysis & Comments

While the primary analysis and experiments of our project are thoroughly detailed above, we also undertook a series of exploratory analyses to delve deeper into our findings. These additional investigations were prompted by specific hypotheses and our innate curiosity about the data. Although insightful, we have chosen not to include these explorations in the main section of our report. This decision is based on the fact that these analyses, while valuable for generating further questions and insights, may not adhere strictly to the rigorous standards of precision and comprehensiveness expected of formal analysis. As such, they should be viewed as preliminary findings that enhance our understanding but require further validation.

(1) Low Accuracy

As observed in the results in Table 2, the models' accuracies were relatively low, compared to the accuracy in our validation model of 81.38%. We hypothesize that this was influenced by the manner in which the test datasets of Messidor and IDRID in our experiments were constructed, coupled with the relatively small number of observations within the datasets. Our experimental setup involved an "unnatural" sampling of the data, which did not reflect the actual distribution found in the datasets. For instance, category 1, which had very few observations in both datasets, was sampled to have an equal number of observations as the other categories in our test dataset. This likely caused the model to predict this category inaccurately. Moreover, with only 24 images in our test dataset, each image influenced the accuracy score by approximately 4.17%, leading to a lower overall accuracy. Since the model architecture was validated using the DDR dataset and the main focus of our experiment was on comparing the relative accuracy of the model to the baseline, we decided to proceed with our main experiments as planned. However, to further explore this matter, we conducted a test using standard sampling method on Messidor raw images, achieving a higher accuracy of around 67% (compared to approximately 58% in Figure 14). This confirmed that one factor contributing to the lower accuracy of our experiment results was the method by which the test dataset was created.

(2) Alternate Evaluation Metric

Since this was a multi-class classification problem, accuracy was a suitable metric, but we also considered using a continuous evaluation metric such as Mean Squared Error (MSE). This consideration arose because Diabetic Retinopathy classification is ordinal and the decision boundaries can be ambiguous, even for human experts. Unlike accuracy, which categorically classifies predictions as true or false, continuous outcome evaluation metrics MSE considers the proximity of the predicted value to the actual value, capturing the degree of error in predictions.

For example, consider the following predictions:
- Actual Category: 4
- Baseline Model Prediction: 2
- Pre-processed Image Model Prediction: 3

Using the accuracy metric, both predictions from the baseline and the pre-processed model would be evaluated equally. However, using MSE, the pre-processed model would score better, reflecting a more accurate prediction closer to the actual category.

To test this hypothesis, we conducted an experiment using the raw-resized Messidor images and the color normalized Messidor images, keeping the data and the model unchanged. The accuracy results as observed in table 2, showed both models scored the same accuracy of 58.33%. However, MSE rated the raw model at 1.04 and the CN-processed model at 1.00, suggesting a slight improvement with preprocessing. These findings indicate that it may be beneficial to explore alternative metrics like MSE to evaluate model performance more comprehensively, even if such metrics are not standard practice for evaluating classification models.

Conclusion

Our project is motivated by the urgent need for effective diabetic retinopathy (DR) detection methods, with the potential to significantly impact global healthcare by improving early screening and intervention for this leading cause of blindness. Leveraging the Messidor and IDRID datasets, we employed the EfficientNetB0 model with a multi-step preprocessing approach including luminosity processing (IENR), cropping (Crop), and color normalization (CN) to enhance model generalization and robustness.

Our study revealed several key findings. The application of luminosity processing did not consistently improve model performance, potentially due to its smoothing effect on crucial features like micro-aneurysms, leading to suboptimal results, particularly in multi-class classification scenarios. Conversely, cropping emerged as a beneficial technique, directing the model's focus towards relevant image features and removing extraneous information. This approach either maintained or improved performance, demonstrating its effectiveness in enhancing predictive capabilities.

When considering color normalization, the results were nuanced. When applied alone, color normalization did not show significant improvements for the Messidor dataset. However, in combination with cropping, it reduced disparities between camera profiles, resulting in improved overall performance. For the Messidor dataset, our baseline model achieved an accuracy of 58.3%, while the model with combined preprocessing (Crop + Color Normalization) achieved an accuracy of 70.8%. On the IDRID dataset, the baseline accuracy was 62.5%, and the model with preprocessing achieved a similar performance.

These results highlight the role of preprocessing in improving model performance, with accuracy as a key evaluation metric. The combination of cropping and color normalization proved particularly effective, significantly enhancing the model's ability to generalize across diverse datasets and camera sources. This approach, evidenced by the notable increase in accuracy from baseline models, suggests its potential for broader applications in DR detection and other medical image analysis tasks.

Looking ahead, these findings suggest avenues for future research. Refinements to preprocessing techniques, exploration of additional methods for model generalization, and investigation of model transferability across datasets are potential areas to explore. Future research could specifically examine the effects of image enhancement and noise removal as separate preprocessing techniques and explore using different methods for performing them. Additionally, employing diverse performance metrics to assess the impact of image preprocessing on diabetic retinopathy (DR) classification would be beneficial. Similar experiments can be performed on different models to see if our findings apply to other types of models as well. These efforts aim to contribute to the development of more effective and robust DR detection systems, with the ultimate goal of benefiting patients worldwide.

Roles

Revanth:
- Managed the Messidor dataset.
- Handled cropping and resizing tasks.
- Implemented Random Forest, Logistic Regression, and ResNet models during the exploratory phase.
- Handled data storage and flow from and to amazon S3
- Created optimized and dynamic notebooks for execution of models
- Collaborated with Tina for EfficientNetB0

Dani:
- Managed the IDRID & DDR datasets.
- Collaborated with Tina on data augmentation.
- Explored alternative normalization methods.
- Collaborated on the CNN512 model, which was explored as an exploratory phase.

Tina:
- Led the development of the EfficientNetB0 model.

- Worked with Afraa on image enhancement and noise removal.
- Collaborated on color normalization with Afraa and Dani.
- Collaborated on the CNN512 model, which was explored as an exploratory phase.

Afraa:
- Collaborated with Tina on image enhancement and noise removal.
- Collaborated on color normalization with Tina and Dani.
- Collaborated on the CNN512 model, which was explored as an exploratory phase.
- Worked on alternative pre-trained model MobileNetV2 during the exploratory phase.

The entire team collectively engaged in running the experiments, results analysis and compiling the project report, ensuring a comprehensive and collaborative effort.

References

Alyoubi, W. L., Abulkhair, M. F., & Shalash, W. M. (2021). Diabetic Retinopathy FundUs image Classification and lesions Localization System using deep learning. *Sensors*, *21*(11), 3704. https://doi.org/10.3390/s21113704

Alyoubi, W. L., Shalash, W. M., & Abulkhair, M. F. (2020). Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, *20*, 100377. https://doi.org/10.1016/j.imu.2020.100377

Ascher, L. M. (2002). Paradoxical intention. In *Elsevier eBooks* (pp. 331–338). https://doi.org/10.1016/b0-12-343010-0/00161-6

Cinaglia, P., Mirarchi, D., & Veltri, P. (2019). Information retrieval in life sciences. In *Elsevier eBooks* (pp. 1104–1108). https://doi.org/10.1016/b978-0-12-809633-8.20390-7

Dhanya, V. G., Subeesh, A., Kushwaha, N. L., Vishwakarma, D. K., Kumar, T. N., Ritika, G. S. S., & Singh, A. (2022). Deep learning based computer vision approaches for smart agricultural applications. *Artificial Intelligence in Agriculture*, *6*, 211–229. https://doi.org/10.1016/j.aiia.2022.09.007

Diabetic Retinopathy Detection | Kaggle. (n.d.). https://www.kaggle.com/c/diabetic-retinopathy-detection/overview

Foracchia, M., Grisan, E., & Ruggeri, A. (2005). Luminosity and contrast normalization in retinal images. *Medical Image Analysis*, *9*(3), 179–190. https://doi.org/10.1016/j.media.2004.07.001

Güler, Ü., Tufan, T. B., Chakravarti, A., Jin, Y., & Ghovanloo, M. (2023). Implantable and wearable sensors for assistive technologies. In *Elsevier eBooks* (pp. 449–473). https://doi.org/10.1016/b978-0-12-822548-6.00072-8

Kansu, T. (2010). Neuro-ophthalmology of Meningiomas. In *Elsevier eBooks* (pp. 177–189). https://doi.org/10.1016/b978-1-4160-5654-6.00012-x

Kotu, V., & Deshpande, B. (2019). Deep learning. In *Elsevier eBooks* (pp. 307–342). https://doi.org/10.1016/b978-0-12-814761-0.00010-1

Lee, R., Wong, T. Y., & Sabanayagam, C. (2015). Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye And Vision*, *2*(1). https://doi.org/10.1186/s40662-015-0026-2

Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., & Kang, H. (2019). Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, *501*, 511–522. https://doi.org/10.1016/j.ins.2019.06.011

Maffre, G. P. G. G. B. L. J. R. D. E. M. F. a. D. H. (2022, May 24). *Messidor - ADCIS*. ADCIS. https://www.adcis.net/en/third-party/messidor/

Oh, K., Kang, H. M., Leem, D., Lee, H., Seo, K. Y., & Yoon, S. (2021). Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-81539-3

Pires, R., Avila, S., Wainer, J., Valle, E., Abràmoff, M. D., & Rocha, A. (2019). A data-driven approach to referable diabetic retinopathy detection. *Artificial Intelligence in Medicine*, *96*, 93–106. https://doi.org/10.1016/j.artmed.2019.03.009

Porwal, P. (2019, November 12). *Indian Diabetic Retinopathy Image Dataset (IDRID)*. IEEE DataPort. https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid

Tan, M., & Le, Q. (2019, May 24). *EfficientNet: Rethinking model scaling for convolutional neural networks*. PMLR. https://proceedings.mlr.press/v97/tan19a.html?ref=jina-ai-gmbh.ghost.io

Vujosevic, S., Aldington, S. J., Silva, P., Hernández, C., Scanlon, P. H., Pető, T., & Simó, R. (2020). Screening for diabetic retinopathy: new perspectives and challenges. ˜*the* œ*Lancet. Diabetes & Endocrinology*, *8*(4), 337–347. https://doi.org/10.1016/s2213-8587(19)30411-5

World Health Organization: WHO. (2019, May 13). *Diabetes*. https://www.who.int/health-topics/diabetes#tab=tab_1

Zochodne, D. W., & Toth, C. (2014). Diabetes and the nervous system. In *Elsevier eBooks* (pp. 351–368). https://doi.org/10.1016/b978-0-12-407710-2.00019-9

Appendix

Appendix 1: Reference images for the classification scales used for Messidor and IDRID

DR Stages	Details	Number	Label
Healthy	Zero abnormalities	548	Normal
Mild NPDR	Microaneurysms	152	Stage 1
Moderate NPDR Severe NPDR	Few microaneurysms Venorous beading + Intraretinal microvascular abnormality	246	Stage 2
PDR	Vitreous/Pre-retinal hemorrhage	254	Stage 3

Image Classification Scale for Messidor

DR Severity Level	Lesions
No DR	No lesions.
Mild DR	MA only.
Moderate DR	More than just MA but less than severe DR.
Severe DR	Any of the following: more than 20 intraretinal HM in each of 4 quadrants; definite venous beading in 2+ quadrants; Prominent intraretinal microvascular abnormalities in 1+ quadrant and no signs of proliferative DR.
Proliferative DR	One or more of the following: neovascularization, pre-retinal HM.

Image Classification Scale for IDRID

Appendix 2: Table Summarizing the Specific Hyperparameters Used for Each Type of Data Augmentation Transformation.

Transformation Type	Description
Rotation	Rotate the image randomly between (−35°, 35°).
Flipping	Horizontal and vertical flip for the images.
Shearing	Randomly shear images with an angle between −15° and 15°.
Translation	Randomly with a shift between −10% and 10% of pixels.
Brightness Range	Randomly darken or brighten the image. Values less than 1.0 darken the image, whereas values

	larger than 1.0 brighten the image. The used values range from 0.25 to 1.25.

Appendix 3: Hyperparameter Configuration for EfficientNetB0

Configuration	**Values**
Optimizer	SGD
Momentum	0.9
Max Learning rate	1×10^−2
Base Learning rate	1×10^−4
Mode	triangular
Class weight	auto
Augmentation	20 times
Dropout	0.5